DOCUMENT RESUME

ED 246 114 TM 840 395

TITLE A Multimethod Approach for Establishing the Passing Score on the Alabama High School Graduation Examination: A Symposium Presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, Louisiana, April 24-26, 1984).

PUB DATE 25 Apr 84

NOTE 58p.

PUB TYPE Collected Works - Conference Proceedings (021) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.

DESCRIPTORS *Cutting Scores; Educational Policy; Graduation Requirements; High Schools; Measurement Techniques; *Minimum Competency Testing; *State Programs; Teacher Attitudes; Testing Programs

IDENTIFIERS *Alabama High School Graduation Examination; Angoff Methods; *Standard Setting

ABSTRACT

These five papers describe the methods, results, and implications of a multimethod approach to the establishment of standards for the three subject areas of the Alabama High School Graduation Examination (AHSGE). James E. McLean introduces the problems and methods involved by identifying three general approaches to scoring procedures found in the literature. These approaches were labeled theoretical, judgmental, and empirical. Robert E. Lockwood discusses the theoretical approach (establishing a passing score independent of exam contents or the examinees' characteristics) using Jensen's probability procedure. Rodney W. Roth explains the judgmental approach: having experts make judgments about test items and objectives. Glennelle Halpin and D. Joyce Steele present the empirical approach: establishing a passing score by evaluating the actual performance of examinees on the examination. Anne C. Hess discusses the policy implications for the standard setting process. During the fall of 1983, the AHSGE was administered statewide for the first time with positive feedback from all involved and a positive impact on improving the quality of education in Alabama. (ES)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED246114

A SYMPOSIUM

A Multimethod Approach for Establishing the Passing Score on the Alabama High School Graduation Examination

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Individual Presentations:

Introduction to the Problem and Methods

James E. McLean
The University of Alabama

The Theoretical Approach: Method and Results

Robert E. Lockwood
Alabama State Department of Education

The Judgmental Approach: Method and Results

Rodney W. Roth
The University of Alabama

The Empirical Approach: Method and Results

Glennelle Halpin
Auburn University

and

D. Joyce Steele
The University of Alabama

Policy Implications of the Standard Setting Process

Anne C. Hess
Alabama State Department of Education

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
McLean, Lockwood, Roth, Halpin & Hess
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A SYMPOSIUM PRESENTED
at the
ANNUAL MEETING
of the
NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION
New Orleans, Louisiana
April 25, 1984

TM 810 355

TABLE OF CONTENTS

Paper	Page
Introduction to the Problem and Methods ..... Author: James E. McLean	1
References	6
The Theoretical Approach: Method and Results ..... Author: Robert E. Lockwood	7
Reference	13
The Judgmental Approach: Method and Results ..... Author: Rodney W. Roth	14
Introduction	15
Rationale for the Angoff Method	15
Data Collection	17
Results	18
References	26
The Empirical Approach: Method and Results ..... Authors: Glennelle Halpin and D. Joyce Steele	27
Method	30
Results	32
References	41
Policy Implications of the Standard Setting Process ..... Author: Anne Hess	42
Reference	48

Discussants: Lawrence H. Cross, Virginia Polytechnic Institute and State University
Stanley Bernknopf, Georgia State Department of Education

A SYMPOSIUM

A MULTIMETHOD APPROACH FOR ESTABLISHING THE PASSING SCORE ON THE ALABAMA HIGH SCHOOL GRADUATION EXAMINATION

Introduction to the Problem and Methods

James E. McLean

The University of Alabama

A Symposium Presented
at the
Annual Meeting
of the
National Council on Measurement in Education
New Orleans, Louisiana
April 25, 1984

INTRODUCTION TO THE PROBLEM AND METHODS

The Alabama High School Graduation Examination (AHSGE) is a product of several Alabama State Board of Education decisions beginning in 1977. On April 27, 1977, the Alabama State Board of Education adopted a resolution which mandated that the State Superintendent of Education appoint a committee to develop minimum competencies for high school graduation and to establish plans for measuring those skills. The Basic Competency Education (BCE) program which grew out of this resolution has as its goal that all Alabama students should have the opportunity to learn the basic skills necessary to earn a credible high school diploma. Following the identification of the standards and competencies at grades 3, 6, and 9, the Basic Competency Tests were developed. The Basic Competency Tests, designed as checkpoints for assessing students' acquisition of these basic skills, have been administered state-wide since 1980.

On May 19, 1981, the Alabama State Board of Education approved new requirements for graduation from high school which mandate that students pass a minimum competency examination in addition to earning the required Carnegie units. In its February 10, 1983, resolution the State Board further specified that the AHSGE be the examination all students must pass to receive a high school diploma.

The AHSGE is composed of three subject tests: Reading, Language, and Mathematics. A student must pass all three subject tests to qualify for a high school diploma. The examination measures the skills identified in the Minimum Standards and Competencies for Alabama Schools booklet and further described in Reading, Language, Mathematics Item Specifications for the Alabama High School Graduation Examination. Copies of these publications have been distributed to all school systems in Alabama.

The items for the AHSGE were pilot tested with a random sample of high school juniors two times. These pilot tests occurred in October, 1982 and April, 1983. The first administration of the AHSGE was in October, 1983. The examination is to be given to 11th grade students each year and they must pass each subject test before the end of their 12th grade year in order to receive a high school diploma. Thus, they will have at least four opportunities to pass the examination.

During the 1982-83 school year, a technical advisory committee was established to address the technical concerns of implementing the AHSGE which is a mandatory requirement for graduation beginning with the class of 1985. The primary function of this Committee was the design, implementation, and interpretation of a study to set the passing scores on the three subject-matter tests (Reading, Mathematics, and Language) comprising the AHSGE. The purpose of this symposium is to present the methods, results, and implications of the methods selected for study by the Committee.

A review of the literature on standard setting procedures revealed that virtually all procedures could be classified in one of three major frameworks based on the source of data under consideration. These frameworks were labeled theoretical, judgmental, and empirical. Procedures classified as theoretical establish a passing score independent of examination content and examinee characteristics. Procedures classified as judgmental establish a passing score by estimating the performance of a hypothetical minimally competent examinee on the examination being used. Finally, procedures classified as empirical establish a passing score by evaluating the actual performance of examinees on the examination. Since the committee agreed that no one procedure within a particular framework addressed all of their concerns, the decision was made to implement one or two procedures from each framework.

The procedure used from the theoretical framework was one proposed by Jensen (1978). This procedure uses the chance score distribution to establish criteria for acceptable performance. Jensen (1978) used probability theory to define a threshold passing score that differentiated between test scores resulting from knowledge of test content and test scores resulting from random guessing. The passing score is defined as the minimum score for which there is a 99% chance that the student who obtains that score does not belong to the distribution of scores obtained by random guessing. The passing score is determined probabilistically from the number of items comprising the test and the number of choices per item.

The judgmental procedure selected for this study was similar to that proposed by Angoff (1971) with two major variations. The first variation was that judges estimated performance for each competency rather than for each item. The second variation was that judges were given two opportunities to judge the difficulty levels of the competencies. Before the second opportunity, the judges were provided with the average item difficulties obtained during the pilot testing for the items in each competency.

Contrasting groups and borderline group methods (Livingston & Zieky, 1982; Zieky & Livingston, 1977) were used from the empirical framework. Respective teachers or counselors most familiar with the reading, mathematics, and language skills of the students sampled in a pilot study were asked: "In your opinion will this student have a sufficient knowledge of [reading, mathematics, or language] to graduate from high school?" Those students for whom the answer was "yes" were designated the masters group, and those for whom the answer was "no" were the nonmasters. A borderline group was formed with those students for whom the answer was "It is questionable."

Both the judgmental and empirical procedures require judgments. For those methods classified as judgmental, judgments are made about the relative

difficulty of the examination (and its content) for a particular group of examinees. Methods described as empirical require judgments about the mastery level of particular examinees prior to test administration. Thus, the frameworks reflect estimated examinee performance (judgmental) and actual performance (empirical).

Since each method identified offered distinct advantages, it was felt that a combination of methods would maximize the validity of the standard setting process. For example, the results from the Jensen method were selected to be the minimum pass rate acceptable. That is, if procedures from the other two frameworks produced pass rates lower than those of the Jensen procedure, the pass rates from the Jensen procedure would be used. Described in the other presentations in this symposium are the implementation, results, and implications of the multimethod solution used for establishing passing scores on the three subject tests (reading, mathematics, and language) comprising the AHSGE.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement. Washington, DC: American Council on Education.

Jensen, O. (1978). Some logical procedures for setting passing points in the absence of direct criterion references. (Internal Document). Princeton, NJ: Educational Testing Service.

Livingston, S. A. & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Zieky, M. J. & Livingston, S. A. (1977). Manual for setting standards on the basic skills assessment tests. Princeton, NJ: Educational Testing Service.

A SYMPOSIUM

A MULTIMETHOD APPROACH FOR ESTABLISHING THE PASSING SCORE ON THE ALABAMA HIGH SCHOOL GRADUATION EXAMINATION

The Theoretical Approach: Method and Results

Robert E. Lockwood

Alabama State Department of Education

A Symposium Presented
at the
Annual Meeting
of the
National Council on Measurement in Education
New Orleans, Louisiana
April 25, 1984

THE THEORETICAL APPROACH: METHOD AND RESULTS

The review of standard setting procedures showed that most procedures could be classified as one of three general types depending on the type of data under consideration. This paper focuses on procedures that can be classified as theoretical. A theoretical procedure is one where the passing score is established independently of the examination content and specific characteristics of the examinees. Thus, a procedure that is based solely on probability theory is classified as a theoretical procedure. One such procedure, proposed by Jensen (1978), was selected as the procedure to be used in conjunction with the judgmental and empirical procedures.

Jensen's procedure uses probability theory to define a threshold passing score that differentiates between test scores resulting from knowledge of test content and test scores resulting from random guessing. Jensen developed the procedure for use in employment testing where the interests of both applicants and employers must be considered. The acknowledgment that both institutions and individuals should be protected when establishing a passing score is appropriate in situations where passing a graduation examination is required to receive a high school diploma.

One of the purposes for requiring students to pass the Alabama High School Graduation Examination (AHSGE) is to renew the credibility of the high school diploma. This renewed credibility is both for students as well as for those who must evaluate the value of the diploma. Consequently, the passing score must assure those who evaluate the meaning of a high school diploma that the possession of a diploma signifies a meaningful level of knowledge in reading, mathematics, and language. At the same time, however, the passing score must also be set so that students who possess the required amounts of knowledge have a fair opportunity to pass the examination. The procedure

proposed by Jensen (1978) acknowledges this dual responsibility and allows for shifts in the passing score depending upon the consequences of certifying or passing less-than-qualified examinees.

The first step in establishing a passing score using Jensen's (1978) procedure is to determine the consequences of passing less-than-qualified examinees. After this determination is made, the passing score is calculated using the chance score distribution. Jensen (1978) provides the following rationale for his theoretical procedure:

> If a perfectly internally consistent, n-item, minimum competency test is scored on the number of right answers basis, the following will be true:
>
> 1. The possible test scores are zero and n.
> 2. Each obtained test score is a true knowledge (T) score.
> 3. The minimum passing point (lowest test score which separates the "know nothings" from the "know-it-alls") is the lowest test score greater than zero; Min pp = 1.
> 4. The maximum passing point equals n and is identical in effect to the Min pp, regardless of whether n is 1 or 1000.
>
> If a fallible n-item test with an average of k choices per item is administered to a large group and then is scored on the number of right answers basis, the following will be true:
>
> 1. Fifty percent of those with a T score of zero (50% of those randomly marking answers to all items) will obtain a test score equal to or greater than n/k.
> 2. One percent of those with a T score of zero (1% of those randomly marking answers to all items) will obtain a test score equal to or greater than $np + 2.326 \sqrt{npq}$ (where $p = 1/k$ and $q = 1 - p$).
> 3. There is no test score that separates all of those that should fail--that separates all those with a T score less than or equal to X from all those with a T score greater than X. (p.1)

The passing score defined as $np + 2.326 \sqrt{npq}$ is appropriate in situations where the consequences of allowing less-than-qualified examinees to pass are relatively minor (defined as low risk by Jensen). As the consequences of error increase, the passing score is raised, still based on the chance score distribution, until there is a 99% probability that the examinee's true score is above the maximum score achieved by 99% of those randomly guessing. The shifting of the passing score depending upon associated risk is shown in Figure 1.

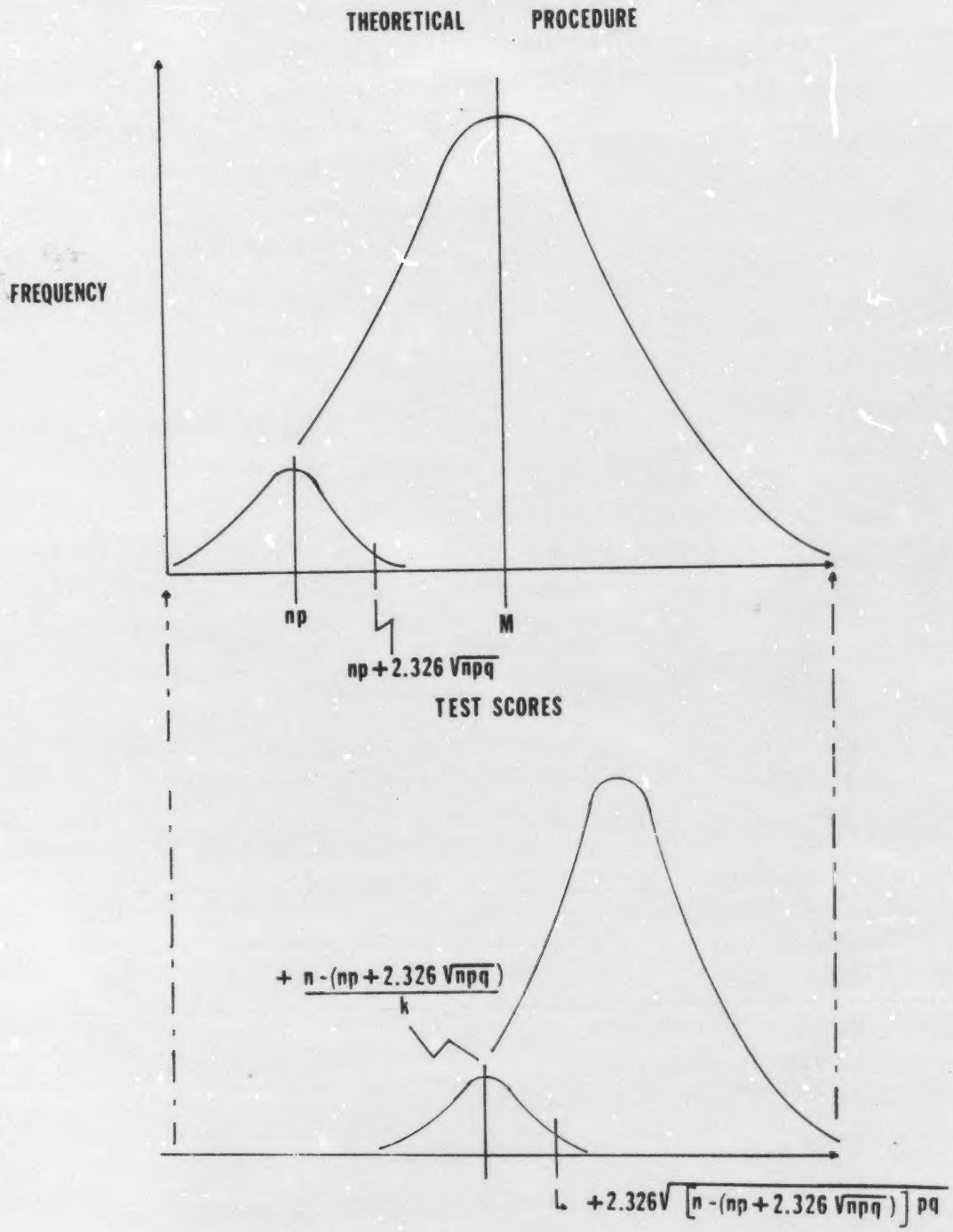

Figure 1: Theoretically Derived Passing Scores

The score labeled np in Figure 1 represents the mean score for the distribution of those randomly guessing on all questions. The score 2.326 standard deviations above that mean represents the 99% confidence level that the true score is greater than zero. This is the passing score for low-risk situations. The lower curve in Figure 1 shows the shifts in the passing score that would occur as a result of an increase in the consequences of error.

For all the items remaining in the test, after subtracting the low-risk passing score, the mean score for the distribution of those randomly guessing on those $n - (np + 2.326\sqrt{npq})$ items is calculated. There is 50% probability that a person obtaining this score has a true score equal to or greater than the test score obtained by 99% of those randomly marking answers to all n items. This represents the next highest passing score (defined as significant risk by Jensen). The final shift yields a score where there is 99% probability that a person obtaining that score has a true score equal to or greater than the test score obtained by 99% of those randomly marking answers to all n items. This is the highest passing score proposed by Jensen and represents what he terms high risk.

Since one of the purposes for the graduation examination is to renew the credibility of the high school diploma and because students have at least four opportunities to pass the examination, it was decided to use the highest passing score proposed by Jensen. Also, since it was recognized that this passing score only represents a specific degree of confidence that the individual is not randomly guessing on all items and because the decision had been made to use procedures from each of the three methodological frameworks, it was decided that the theoretically derived passing score would serve as a benchmark for procedures from the judgmental and empirical frameworks. That is, in the event that the judgmental and empirical procedures produced passing

scores below that defined using the Jensen (1978) procedure, the theoretically defined passing score would be used.

The results of applying the standard setting procedure from the theoretical framework in the three subject area tests comprising the AHSGE are presented in Table 1.

Table 1

Results of Theoretical Standard Setting Procedure

Subject Area	No. of Items	No. of Options Per Item	Derived Passing Score
Reading	80	4	49 (61%)
Mathematics	95	4	57 (60%)
Language	119	4	70 (59%)

Thus, passing scores of 49, 57, and 70 for Reading, Mathematics, and Language respectively became the minimum passing scores considered.

Reference

Jensen, O. (1978). Some logical procedures for setting passing points in the absence of direct criterion references. Unpublished manuscript, Educational Testing Services, Princeton, NJ.

A SYMPOSIUM

A MULTIMETHOD APPROACH FOR ESTABLISHING THE PASSING SCORE ON THE ALABAMA HIGH SCHOOL GRADUATION EXAMINATION

The Judgmental Approach: Method and Results

Rodney W. Roth

The University of Alabama

A Symposium Presented
at the
Annual Meeting
of the
National Council on Measurement in Education
New Orleans, Louisiana
April 25, 1984

THE JUDGMENTAL APPROACH: METHOD AND RESULTS

Introduction

This paper, as part of a symposium entitled "A Multimethod Approach for Establishing the Passing Score on the Alabama High School Graduation Examination," presents the judgmental approach. This was one of several approaches used to derive the passing score for the Alabama High School Graduation Examination (AHSGE).

A Technical Advisory Committee (TAC), composed of persons from postsecondary institutions and the Alabama State Department of Education, was established to design and implement a study to "determine" the passing scores on the three subject-matter tests comprising the AHSGE. The derived scores were recommendations from the TAC to the Alabama State Board of Education. This body legally set the three passing scores.

Rationale for the Angoff Method

The TAC conducted a literature review of the various judgmental procedures which could be used for determining a passing score. In addition, several members (Lockwood, 1979; Halpin, 1980; Roth, 1982) of the TAC had been involved with research and applications of various standard setting procedures. Some of the research studies that led the TAC to select the modified Angoff method for the judgmental procedure are presented in this section.

One study (Cross, Impara, & Frary, 1983) compared Angoff, Nedelsky, and Jaeger standard setting procedures for the NTE area examinations in mathematics and elementary education. The authors concluded:

1. The Angoff procedure yielded standards that, while quite stringent, were more realistic than those associated with the Jaeger method and more stable across examinations than the Nedelsky method.

2. The reliability and validity indices were enhanced most dramatically with the introduction of the item difficulty values during the second session. This was particularly true for the Angoff method which during the second session resulted in knowledge-estimation ratings that were unsurpassed with respect to these psychometric properties.

3. When asked at the end of the review sessions to express how much confidence they placed in the procedures used, those using the Angoff method expressed greater confidence in their knowledge estimates and greater confidence in the standards resulting from these judgments, on the average, than those using the other methods.

A study (Halpin, 1980) using ten different standard setting procedures for the Missouri College English Test produced a wide range of passing scores. The range for the ten procedures was from the 9th percentile to the 88th percentile. The average percentile of these 10 procedures was the 36th percentile. This was also the percentile produced by using the Angoff method.

In a related study, Halpin & Halpin (1981) determined the reliability and validity of the Ebel, Nedelsky, and Angoff methods of standard setting. They found, based on the internal consistency of the judges' ratings, the intercorrelations among the three methods, and the correlations between item ratings, and the external criteria of actual item difficulty that the Ebel and Angoff standard setting methods, in contrast to the Nedelsky method, are the most reliable and valid of the three.

The TAC did discuss the possibility of using the Ebel method. It was decided, however, that all of the competencies on the AHSGE are minimum and have equal importance. Thus, the Ebel procedure was eliminated because the judges are asked to rate the importance of each objective or item.

Data Collection

A primary characteristic of all judgmental standard setting procedures is that persons make judgments about the test items or test objectives. It is extremely important that the persons serving as judges have the appropriate background and experience to render expert judgment. The TAC thoroughly discussed the judge selection problem. The TAC decided that the most appropriate set of judges would be Alabama classroom teachers who had taught for several years. In addition, the teachers would be ones who taught remedial high school classes or ninth and tenth grade general classes.

The final set of judges consisted of 11 math teachers and 23 English teachers. Eleven of the English teachers judged the reading test, and the remaining 12 English teachers judged the language test.

The 34 judges met at one site for training on defining minimally competent examinees and the specific procedures for judging the test objectives. At the beginning of the training session, the judges were given the following definition of a minimally competent student. It was:

> A student who can complete the required Carnegie units and has a minimum knowledge of [reading or math or language] to graduate from high school.

The judges were asked to visualize a student who they believed met the definition of a minimally competent student. In other words, the student would be one perceived who was just above the "line" that divided students into two categories: those who possessed minimum knowledge to earn a high school diploma and those who should not receive a diploma because of not possessing enough minimal knowledge.

The judges then started the process which would derive a passing score based on the judgmental procedure. The judges were given the AHSGE competencies for their subject expertise and a set of test items associated

with each competency. The judges were instructed to read the competency and to study the test items measuring the competency. The judges were asked to estimate the proportion or percent of items in each competency the minimally competent student they had visualized should answer correctly.

The judges were given time to complete the estimates for each competency on their subject matter test. Upon completion, each judge was presented with a list of average item difficulties for each competency which were determined during the April pilot test. After an explanation of their meaning and some cautions about the possible undue influence of these data, the judges were asked to review their judgments and modify them if they felt that modifications were appropriate.

Results

The judgments made by the judges during the second round (average item difficulties were presented) were used to determine the passing score from the judgmental procedure. The first step in the analysis was to determine an average percentage expected by the judges for each competency. This average was multiplied by the number of items measuring the competency on the test. The sum of the expected items for each competency gives the total number of items needed to pass the test.

The derived passing score for reading was 60 items or 75%. For math, the number of items was 62 or 65%, and for language, the number of items was 96 or 81%. The results for the three tests are in Tables 1-3.

The reliabilities (coefficient alphas) were computed using the second round judgments. They were .93, .98, and .88 for reading, math, and language respectively.

Table 1

Passing Scores Set Using Judgmental Procedure for Reading

	Competency	Number of Items Measuring Competency	Average Percentage Expected by Judges	Number of Items Based on Expected Percentage
I-1, 2; IV-13	Associate selected written words and phrases which occur in daily life situations with their literal meanings; recognize and use various forms and application blanks used in everyday life.	8	76.36	6.1088
II-1	Determine word meaning using specific context clues	5	72.73	3.6365
III-1	Determine word meanings from a knowledge of prefixes and suffixes.	4	68.64	2.7456
III-2	Recognize common abbreviations and symbols.	4	83.18	3.3272
IV-1	Use alphabetical organization to locate information needed.	6	76.36	4.5816
IV-3	Select appropriate reference sources in a given situation.	4	79.09	3.1636
IV-4	Know the uses of the different sections of a book.	4	73.18	2.9272
IV-5	Know the uses of the different sections of the newspaper.	4	80.45	3.2180
IV-6	Determine the main idea in a passage.	5	70.00	3.5000
IV-7	Find details in a written passage.	5	76.36	3.8180
IV-8	Use maps, graphs, charts, and tables.	6	74.09	4.4454
IV-9, 10, 11	Determine main idea and detail from selected materials on government, law, money management, health, and nutrition.	9	70.91	6.3819

Table 1 Continued

	Competency	Number of Items Measuring Competency	Average Percentage Expected by Judges	Number of Items Based on Expected Percentage
IV-14	Read for information from materials such as advertisements, yellow pages, labels, directories, and catalogs.	6	81.55	4.8930
V-1	Follow directions and instructions.	6	79.05	4.7454
VI-1	Distinguish fact from fiction.	4	71.36	2.8544
Total		80	75.43	60.3466

Table 2

Passing Scores Set Using Judgmental Procedures for Mathematics

	Competency	Number of Items Measuring Competency	Average Percentage Expected by Judges	Number of Items Based on Expected Percentage
I-1	Read and write numerals and word names for numbers	3	76.00	2.2800
I-2	Compare decimals; compare fractions with unlike denominators	2	57.73	1.1546
I-3	Add, subtract, multiply, and divide whole numbers.	6	82.73	4.9638
I-4	Add, subtract, multiply, and divide decimals.	7	70.09	4.9063
I-5	Add, subtract, and multiply fractions.	6	61.82	3.7092
I-6	Change percents to decimals and decimals to percents.	3	65.55	1.9665
I-7	Determine discounts, down payments, commissions, taxes, ratios, and percents	4	59.27	2.3708
I-8	Convert fractions to decimals and decimals to fractions.	4	58.36	2.3344
I-9	Find averages of whole numbers and decimal numbers.	3	69.18	2.0754
I-10	Change units of measure.	4	51.18	2.0472
II-1	Solve problems involving time.	3	63.18	1.8954
II-2	Measure distance by use of a ruler.	4	60.00	2.4000
II-3	Measure capacities.	3	65.55	1.9665
II-4	Find perimeters.	4	69.09	2.7636
II-5	Find rectangular areas.	4	53.64	2.1456
III-1	Read and interpret graphs.	2	63.18	1.2636

Table 2 Continued

	Competency	Number of Items Measuring Competency	Average Percentage Expected by Judges	Number of Items Based on Expected Percentage
III-2	Read and interpret tables and charts.	2	64.27	1.2854
III-3	Interpret scale drawings.	3	62.27	1.8681
IV-1	Make approximations by rounding numbers.	3	66.00	1.9800
V-1, 3	Recognize simple plane geometric figures, parallel and perpendicular lines.	4	77.73	3.1092
VI-1	Determine change.	2	70.91	1.4182
VI-2	Use comparative shopping skills	2	63.91	1.2782
VI-3	Compare the cost of cash and credit purchases.	2	57.73	1.1546
VI-4	Allocate time and money.	3	61.91	1.8573
VI-5	Complete checks and money orders.	3	67.55	2.0265
VI-6	Solve rate, time, distance problems.	4	64.64	1.9392
VI-7	Compute salaries and wages.	4	61.36	2.4544
Total		95	65.29	62.0230

Table 3

Passing Score Set Using Judgmental Procedure for Language

	Competency	Number of Items Measuring Competency	Average Percentage Expected by Judges	Number of Items Based on Expected Percentage
I-1	Observe pronoun-antecedent agreement	9	83.33	7.4997
I-2	Use correct forms of nouns and verbs.	15	80.25	12.4997
II-1	List ideas in sequential order.	5	79.33	3.9665
II-2	Arrange sentences into a logically organized paragraph.	9	75.58	6.8022
II-3	Arrange written directions in logical order.	4	81.92	3.2768
IV-1	Include in a message or request all necessary information (who, what, when, where, how, or why).	4	82.50	3.3000
IV-2	Determine what information is missing from a message, an announcement, or a process explanation; or what information is irrelevant.	4	82.75	3.3100
V-1	Identify proper format of a business letter.	9	77.75	6.9975
V-2	Identify the proper format of a friendly letter (social correspondence).	8	80.92	6.4736
VI-1	Complete a common form such as a driver's license application, change of address form.	9	83.75	7.5375
VII-1	Identify words frequently used in daily activities.	9	86.25	7.7625
VIII-1, 2, 3, 4	Identify question marks, periods, and exclamation points to punctuate sentences.	16	81.50	13.0400
VIII-5	Identify quotation marks to set off a direct quotation.	4	73.33	2.9332

Table 3 Continued

	Competency	Number of Items Measuring Competency	Average Percentage Expected by Judges	Number of Items Based on Expected Percentage
VIII-6	Identify the colon after "the following" and "as follows," the salutation of a business letter, and the hour in numerals.	4	75.00	3.0000
IX-1	Identify the capitalization of the first word in every sentence, the pronoun I, and proper nouns.	10	81.03	96.4290
Total		119	81.03	94.4290

One additional result of interest was the lack of total score difference between round one and round two. These results are presented in Table 4. Using the data from this study, McLean and Lockwood (1983) studied the sources of variability. The sources examined were judges, rounds, competencies, and the various interactions. They found the judge by competency variance component accounted for the highest proportion of variance. The "round" component accounted for an extremely small amount of the variance

Table 4

Estimates of Total Scores by Judges for Rounds 1 and 2

Subject	Mean		Mean difference
	Round 1	Round 2	
Reading	74.758	75.558	.800 NS
Mathematics	64.747	64.734	.013 NS
Language	79.839	80.578	.739 NS

NS--nonsignificant difference using given F-ratio.

References

Cross, L., Frary, R., & Impara, J. (1983). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Paper presented at American Educational Research Association meeting, Montreal, Canada, April.

Halpin, G. (1980). Setting minimum English proficiency standards. Paper presented at Mid-South Educational Research Association meeting, New Orleans, Louisiana, November.

Halpin, G., & Halpin, G. (1981). Reliability and validity of the Ebel, Nedelsky, and Angoff methods of standard setting. Paper presented at Mid-South Educational Research Association meeting, Lexington, Kentucky, November.

Lockwood, R. (1979). Application of generalizability theory to the issue of agreement among raters in setting cutting scores. Paper presented at Mid-South Educational Research Association meeting, Little Rock, Arkansas, November.

McLean, J. E. & Lockwood, R. E. (1983). Sources of error in an Angoff type standard setting process. Paper presented at Mid-South Educational Research Association meeting, Nashville, Tennessee, November.

Roth, R. (1982). The Arkansas National Teacher Examination (NTE) validation and cut-score study. Fayetteville, Arkansas: University of Arkansas, June.

A SYMPOSIUM

A MULTIMETHOD APPROACH FOR ESTABLISHING THE PASSING SCORE ON THE ALABAMA HIGH SCHOOL GRADUATION EXAMINATION

The Empirical Approach: Method and Results

Glennelle Halpin
Auburn University

and

D. Joyce Steele
The University of Alabama

A Symposium Presented
at the
Annual Meeting
of the
National Council on Measurement in Education
New Orleans, Louisiana
April 25, 1984

THE EMPIRICAL APPROACH: METHOD AND RESULTS

In competency testing programs, the identification and use of appropriate procedures for determining minimum passing scores are very important. Those faced with this critical decision are offered a variety of choices (cf. Halpin & Halpin, 1982). Among these choices are methods requiring judgments about the mastery levels of actual test takers rather than the test items. After these test takers have been subdivided into mastery groups, empirical data--scores of these examinees on the tests for which the passing scores are to be set--are then used in determining cut-off points.

It has been suggested that judgments about students are more meaningful and less difficult than judgments of test items (Zieky & Livingston, 1977). Teachers especially are accustomed to separating masters and nonmasters with respect to educational objectives. Therefore, they might more validly identify the student reference groups which form the basis for all subsequent steps in the empirical approach than make judgments about the test performance of hypothetical minimally competent individuals on the test items. Even if not preferable to item judgment methods, the empirical method certainly adds a dimension to standard setting that should be considered. It provides known-groups data against which to check the validity of passing scores set with other methods (Koffler, 1980).

Choosing the empirical approach, practitioners may select the borderline group method, and/or they may use the contrasting groups approach (cf. Berk, 1976; Livingston & Zieky, 1982). With the borderline group method, individuals (for example, teachers) familiar with the competence of students in areas being tested are asked to identify those at or near the borderline with respect to mastery of the competencies of interest. The average

performance of these borderline students so identified on a test of the competencies becomes the passing score.

In the contrasting groups approach, judges (e.g., teachers) choose those students clearly masters or nonmasters of the competencies in question. Any one of several methods may then be used to calculate passing scores using actual test data for these students. Perhaps the simplest of these utilizes a graph of the frequency/score distributions for the masters and nonmasters groups which should be approximately equal in number (Berk, 1976). The point where the two distributions intersect is the optimal cut-off point.

With a non-normal distribution and unequal variances, Koffler (1980) used a quadratic discriminant function (QDF) procedure with ranks replacing student test scores to set contrasting groups cut-off scores. Interestingly, he found that with this method all students would have passed one test regardless of their test scores. Mills (1983) and Poggio, Glasnapp, and Eros (1982) reported similar results.

Livingston and Zieky (1982) suggested yet a third way that contrasting groups passing scores may be set which is based on the probability of mastery at each score level. The passing score is the score at which 50% of the students are masters with the assumption being that at the lower test score levels the proportion unqualified increases. Another important assumption with this method is that the test takers who are being judged are a representative sample of all test takers, i.e., the conditional probability of being judged adequate, given the test taker's test score, is the same in the sample as in the population. Equal numbers of masters and nonmasters would therefore be a biased sample (Livingston, 1982).

Considering our group--a stratified random sample of 11th graders in Alabama--which had an unequal number of masters and nonmasters plus reviewing the assumptions and limitations of various contrasting groups approaches, we

chose to use Livingston and Zieky's method to set contrasting groups passing scores.

Procedures and results follow for both the contrasting groups and the borderline group procedures.

Method

Schools throughout Alabama were ranked according to average performance by their students on the California Achievement Tests. From each of five low to high achievement strata identified, schools were randomly selected to participate in the piloting of two forms of the Alabama High School Graduation Examination (AHSGE) measuring reading, mathematics, and language. Within each of those schools, from 6 to 88 11th grade students were randomly selected to take the examination. Half of the students in each school responded to one pilot form of the test while the other half took the alternate form. (See Table 1 for sample sizes, test lengths, means, standard deviations, reliability estimates, and standard errors of measurement.)

In each school, the teacher or teachers most familiar with each student's respective mastery of the basic competencies being measured by the reading, mathematics, and language tests were asked:

> "In your opinion, will this student have a sufficient knowledge of [reading, mathematics, language] to graduate from high school?"
> Possible answers: (1) Yes, (2) It is questionable, or (3) No.

This information was collected prior to testing, generally from students' mathematics teachers for mathematics and English teachers for language and reading, and coded on the student answer sheets after testing.

Students for whom the answer was "yes" were designated as masters for standard setting purposes. Those for whom the answer was "no" became the nonmasters. The borderline group was made up of the students for whom the answer was, "It is questionable."

Table 1

Summary Statistics for the Pilot Forms of the AHSGE

Subject Form	N	Number of Items	$\bar{X}$	SD	SEM	Cronbach Alpha
Reading						
Form 11	609	80	72.03	10.66	.43	.95
Form 12	610	80	71.37	11.14	.45	.95
Mathematics						
Form 23	610	95	72.77	16.26	.66	.95
Form 24	610	95	73.18	16.11	.65	.96
Language						
Form 35	615	119	99.12	17.54	.71	.96
Form 36	616	119	98.32	18.58	.75	.96

In setting the contrasting groups passing scores using the method described by Livingston and Zieky (1982), test score intervals were established, and the smoothed percent of masters and nonmasters at each interval was determined using the moving averages method of smoothing. Results for reading, mathematics, and language are shown in Tables 2, 3, and 4.

For the reading, mathematics, and language tests the decision rule was to set the passing score at the lower limit of the score interval where 50% of the examinees fell into the mastery category and 50% in the nonmastery category since it was generally deemed equally serious to misclassify either masters or nonmasters. When an exact 50-50 point was not available, the cut-off was determined by finding the interval where the proportions of students classified as masters and nonmasters were closest to the 50-50 point. The higher interval was chosen where the 50-50 point (or a close approximation) occurred twice since, due to a larger number of examinees at the higher point, the results should be more stable.

For the borderline group, means were computed for the reading, mathematics, and language tests. These means became the reading, mathematics, and language passing scores set with the borderline group method.

Results

Passing scores obtained with the borderline and contrasting groups methods are shown in Table 5. With the borderline method these were 64 of 80 items on the reading test (80%), 61 of 95 items on the mathematics test (64%), and 87 of 119 items on the language test (73%). For the contrasting groups approach passing scores were 53 of 80 reading items (66%), 40 of 95 mathematics items (42%), and 60 of 119 language items (50%).

Table 2

Summary of Performance of Mastery and Nonmastery Groups in Reading

Score intervals	Mastery frequency	Nonmastery frequency	Total frequency	Smoothed percent of mastery group	Smoothed percent of nonmastery group
77-80	515	4	519	*	*
73-76	208	6	214	97.63	2.37
69-72	100	10	110	93.26	6.74
65-68	38	9	47	85.71	14.29
61-64	24	8	32	73.53	26.47
57-60	13	10	23	61.90	38.10
53-56	2	6	8	47.73	52.27
49-52	6	7	13	40.00	60.00
45-48	2	2	4	37.50	62.50
41-44	1	6	7	35.00	65.00
37-40	4	5	9	38.10	61.90
33-36	3	2	5	50.00	50.00
29-32	1	1	2	44.44	55.56
25-28	0	2	2	12.50	87.50

Table continues

Table 2 (continued)

Score intervals	Mastery frequency	Nonmastery frequency	Total frequency	Smoothed percent of mastery group	Smoothed percent of nonmastery group
21-24	0	4	4	0.00	100.00
17-20	0	4	4	0.00	100.00
0-16	0	2	2	*	*

*The smoothing method cannot be used to estimate the smoothed percent at the lowest and the highest test score intervals.

Table 3

Summary of Performance of Mastery and Nonmastery Groups in Mathematics

Score intervals	Mastery frequency	Nonmastery frequency	Total frequency	Smoothed percent of mastery group	Smoothed percent of nonmastery group
92-95	69	2	71	*	*
88-91	155	3	158	98.01	1.99
84-87	121	2	123	97.52	2.48
80-83	118	5	123	97.18	2.82
76-79	105	3	108	96.08	3.92
72-75	71	4	75	95.08	4.92
68-71	56	5	61	93.12	6.88
64-67	49	4	53	89.24	10.76
60-63	36	8	44	88.64	11.36
56-59	32	3	35	87.50	12.50
52-55	23	2	25	82.93	17.07
48-51	13	9	22	73.13	26.87
44-47	13	7	20	57.14	42.86
40-43	6	8	14	52.27	47.73
36-39	4	6	10	43.33	56.67
32-35	3	3	6	45.00	55.00

Table continues

Table 3 (continued)

Score intervals	Mastery frequency	Nonmastery frequency	Total frequency	Smoothed percent of mastery group	Smoothed percent of nonmastery group
28-31	2	2	4	40.00	60.00
24-27	1	4	5	25.00	75.00
0-23	0	3	3	*	*

*The smoothing method cannot be used to estimate the smoothed percent at the lowest and highest test score intervals.

Table 4

Summary of Performance of Mastery and Nonmastery Groups in Language

Score intervals	Mastery frequency	Nonmastery frequency	Total frequency	Smoothed percent of mastery group	Smoothed percent of nonmastery group
115-119	169	1	170	*	*
110-114	211	1	212	99.49	0.05
105-109	202	1	203	98.08	2.92
100-104	97	8	105	96.44	3.56
95-99	80	5	85	92.00	8.00
90-94	53	7	60	87.96	12.04
85-89	35	11	46	80.00	20.00
80-84	20	9	29	68.75	31.25
75-79	11	10	21	60.00	40.00
70-74	11	9	20	56.90	43.10
65-69	11	6	17	55.10	44.90
60-64	5	7	12	52.94	47.06
55-59	2	3	5	40.00	60.00
50-54	3	5	8	33.33	66.67
45-49	1	4	5	37.50	62.50
40-44	2	1	3	38.46	61.54

Table continues

Table 4 (Continued)

Score intervals	Mastery frequency	Nonmastery frequency	Total frequency	Smoothed percent of mastery group	Smoothed percent of nonmastery group
35-39	2	3	5	41.67	58.33
30-34	1	3	4	23.08	76.92
0-29	0	4	4	*	*

*The smoothing method cannot be used to estimate the smoothed percent at the lowest and highest test score intervals.

Table 5

Passing Scores Set with Empirical Methods

Test form	Number of items	Contrasting Groups		Borderline Group	
		Passing score	% items	Passing score	% items
Reading	80	53	66	64	80
Mathematics	95	40	42	61	64
Language	119	60	50	87	73

In the final analysis of the tentative passing scores set with the different methods, two problems were found with the contrasting groups approach that resulted in its being dropped from further consideration. First, since the percentage qualified was approximately 50% at two different score intervals on the reading test, there was no clear cut-off point. The higher point was tentatively chosen, but the lower point might have also been selected. Further, the number of students identified as nonmasters was lower on all three tests than the minimum number needed for a stable cut score (cf. Huynh, 1981; Popham, 1978). Consequently, only the cut-off scores from the borderline group method were utilized in selection of final passing scores for the Alabama High School Graduation Examination.

References

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education, 45, 4-9.

Halpin, G., & Halpin, G. (1982). Setting performance standards: A documentation and discussion. Education, 103, 118-128.

Huynh, H. (1981, April). Technical and practical considerations in setting standards. Paper presented at the meetings of the American Educational Research Association and the National Council on Measurement in Education, Los Angeles, CA.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17, 167-178.

Livingston, S. A. (1982, March). Assumptions of standard setting methods. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Mills, C. N. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. Journal of Educational Measurement, 20, 283-292.

Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1982, March). An evaluation of contrasting groups methods for setting standards. Paper presented at the meeting of the American Educational Research Association, New York, NY.

Popham, W. J. (1978). Setting performance standards. Los Angeles: Instructional Objectives Exchange.

Zieky, M. J., & Livingston, S. A. (1977). A manual for setting standards on the Basic Skills Assessment Tests. Princeton, NJ: Educational Testing Service.

A SYMPOSIUM

A MULTIMETHOD APPROACH FOR ESTABLISHING THE PASSING SCORE ON THE ALABAMA HIGH SCHOOL GRADUATION EXAMINATION

Policy Implications of the Standard Setting Process

Anne C. Hess

Alabama State Department of Education

A Symposium Presented
at the
Annual Meeting
of the
National Council on Measurement in Education
New Orleans, Louisiana
April 25, 1984

POLICY IMPLICATIONS OF THE STANDARD SETTING PROCESS

It was the intent of the Alabama State Department of Education and the Technical Advisory Committee that the method for establishing standards for the three subject areas of the Alabama High School Graduation Exam (AHSGE) be technically sound and consistent. Beyond technical concerns, many philosophical and moral issues had to be addressed before procedures could be selected and passing scores determined. The establishment of these scores would have a significant impact on many facets of society.

Other considerations often have been a part of establishing standards for high school graduation examinations across the country. For example, the amount of money available for remediation of those students who fail has influenced the standards. Political implications of the number of students anticipated to fail and potential disparate passing rates for various groups of students are also influential factors. While these may be legitimate concerns, a distinguishing aspect of the standard setting process for the AHSGE was an adherence to psychometric concerns and the absence of political pressure for a specific passing rate. The State Superintendent of Education and all concerned supported the belief that a standard should be set based on what is expected of all students regardless of the number of anticipated failures of monetary considerations. Item pretest data predicting numbers of failures, therefore, were never presented to the Technical Advisory Committee during the standard setting process.

Educators' input and decision making have been an integral part of the development of the Alabama Basic Competency Education program. The determination of the passing scores for the AHSGE supports this philosophy by providing teacher input in two ways—judging the difficulty of items measuring

the competencies and judging the performance level of students. The use of judgment is not, as some would say, merely an arbitrary decision.

> The establishment of a minimum competency level can be done only through the use of informed human judgment. There are no magic statistical formulas or objective guidelines that can be applied easily to that task. If one rejects the right or the responsibility of educators to use informed human judgment in such a decision, one must also turn one's back to the hundreds and thousands of pass-fail decisions that are made in education every day. (Gallagher, 1979)

Human opinion and judgment is, in fact, the basis of our legal process as well as everyday life. As Gallagher also points out, the number of credits required for high school graduation, which has been the criteria for years, is also based on human judgment. The public accepts judgment as a decision process. Our obligation is to be sure those making the decision are qualified to do so. Political leaders must accept standards set by those most qualified without dismissing them for political reasons.

Three procedures described in previous sections of this paper were used to establish the standards for the AHSGE. The passing score established by the Jensen procedure did not come into play since the other procedures resulted in passing scores above the theoretical threshold. The minimum passing scores recommended to the Alabama State Board of Education, therefore, were based on the average of the judgmental (Angoff) and empirical (Borderline Group) methods. The combination resulted in the following cutting scores: 63 of 80 reading items (79%), 62 of 95 mathematics items (65%), and 92 of 119 language items (77%).

The specific score established as the minimum passing score for the AHSGE has varying implications for students, parents, teachers, the curriculum, and employers; affecting each of these segments of the population differently. Employers and lay citizens view a graduation examination as renewing credibility in the diploma. Students perceive themselves either as the

beneficiaries or victims of the process. If they pass, their diplomas are viewed more favorably; but if they fail, they will not receive a diploma. To this point, parents have been supportive of the AHSGE requirement. However, no student has yet been denied a diploma. Teachers not only must consider possible curricular modifications, but they are seen as directly responsible for reteaching skills to students who fail the test and documenting the remediation process.

The AHSGE was administered statewide for the first time during the fall of 1983 and received very positive feedback from students, teachers, the press, and the general public. Students interviewed after taking the AHSGE felt that it was not easy, but, at the same time, believed that a student should not graduate without passing it. Members of the state press were given an opportunity to take the AHSGE under testing conditions similar to those required of the students. They agreed that it was not a pushover, but that it was a fair assessment of basic skills in reading, mathematics, and language.

The results of the examination were well received because passing rates were higher than expected in most school systems. There were, however, some school systems whose lower scores have caused concern to the State Department of Education and local administrators. This concern is reflected in the State Board of Education's mandating that assistance be given to these school systems to identify and rectify the causes of these lower scores. This assistance, which will ultimately strengthen the entire instructional process within these systems, is a direct outgrowth of the AHSGE.

The philosophy of the entire Basic Competency Education program should not be buried by the impact of standard setting and the possible denial of diplomas. The purpose of the Basic Competency Education program was, and still is, to provide an instructional process that assures each student the opportunity to master basic skills in reading, mathematics, and language.

Competency tests are given to students in grades 3, 6, and 9 as checkpoints for the attainment of mastery of the basic skills. The ninth grade test serves as a crucial checkpoint as it contains the same skills that are tested on the AHSGE. Remediation, therefore, can begin in the tenth grade before the graduation examination which is administered to students for the first time in the fall of their eleventh grade with three subsequent opportunities to pass it before the end of the twelfth grade. Performance on the checkpoint tests are not evaluated in terms of pass-fail but rather for diagnostic and remediation purposes. That is, student performance is evaluated as mastery or nonmastery for each of the competencies comprising each of the three subject areas. The intent of the AHSGE is to identify students deficient in the minimum skills and to then provide remediation for those skills, only denying an Alabama High School diploma when a student has failed to attain mastery at the end of the twelfth grade.

Even though, philosophically, the third, sixth, and ninth grade tests were not designed to be pass-fail tests, the fact that the graduation examination has a passing score has created an interest in establishing passing scores in grades 3, 6, and 9 for possible promotion/retention purposes. This request requires a re-evaluation of the purposes and philosophy of the testing at earlier grades.

Testing has become a prominent issue in Alabama since the implementation of the Basic Competency Education program and particularly since the reality of the AHSGE. Teachers and administrators have become more knowledgeable about interpreting test data for competency tests as well as for norm-referenced tests. Results are being used for instructional purposes and school systems have evaluated their curricula to ensure that students are taught what is tested and that tests are satisfactorily assessing a school system's curriculum. An emphasis has also had to be placed on assuring that

minimums do not become maximum expectations. The testing program has caused intense study and critique of "what is" and "what should be" in each school system.

The entire standard setting process was characterized by a desire to improve the quality of education in Alabama. The intent was to assure that students can meet the standard and that passing the examination is evidence of knowledge in each of the subject areas. The reality of setting standards has had a tremendous impact on the state. The fact that they are technically sound, defensible, and void of political adjustments is a source of pride for the State Department of Education.

Reference

Gallagher, J. J. (1979). Minimum competency: The setting of educational standards. Educational Evaluation and Policy Analysis, 1(1), 62-67.